Understanding Affective Content of Music Videos Through Learned Representations

Esra Acar, Frank Hopfgartner, and Sahin Albayrak

DAI Laboratory, Technische Universität Berlin,
Ernst-Reuter-Platz 7, TEL 14, 10587 Berlin, Germany
{name.surname}@tu-berlin.de

Abstract. In consideration of the ever-growing available multimedia data, annotating multimedia content automatically with feeling(s) expected to arise in users is a challenging problem. In order to solve this problem, the emerging research field of video affective analysis aims at exploiting human emotions. In this field where no dominant feature representation has emerged yet, choosing discriminative features for the effective representation of video segments is a key issue in designing video affective content analysis algorithms. Most existing affective content analysis methods either use low-level audio-visual features or generate hand-crafted higher level representations based on these low-level features. In this work, we propose to use deep learning methods, in particular convolutional neural networks (CNNs), in order to learn mid-level representations from automatically extracted low-level features. We exploit the audio and visual modality of videos by employing Mel-Frequency Cepstral Coefficients (MFCC) and color values in the RGB space in order to build higher level audio and visual representations. We use the learned representations for the affective classification of music video clips. We choose multi-class support vector machines (SVMs) for classifying video clips into four affective categories representing the four quadrants of the Valence-Arousal (VA) space. Results on a subset of the DEAP dataset (on 76 music video clips) show that a significant improvement is obtained when higher level representations are used instead of low-level features, for video affective content analysis.

Keywords: Affect Analysis, Learning Feature Representations, Convolutional Neural Network, Support Vector Machine

1 Introduction

Accessing online videos through services such as video on demand has become extremely easy thanks to equipments including DVB set top boxes (terrestrial, cable or satellite), Tablet PCs, high-speed Internet access or digital media-streaming devices. However, among the growing amount of multimedia, finding video content matching the current mood and needs of users is still a challenge. Within this context, video affective analysis is an emerging research field that targets this problem of affective content analysis of videos. The affective content of a video is defined as the intensity (i.e., arousal) and type (i.e., valence) of emotion (both are referred to as affect) that are expected to arise in the user while watching that video [7]. In order to deal with this

challenging affective analysis problem, machine learning methods are mainly used. The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied [2]. Therefore, one key issue in designing video affective content analysis algorithms is the representation of video content as in any pattern recognition task. The common approach for video content representation is either to use low-level audio-visual features or to build hand-crafted higher level representations based on the low-level ones (e.g., [5,8,13,21]). Low-level features have the disadvantage of losing global relations or structure in data, whereas creating hand-crafted higher level representations is time consuming, problem-dependent, and requires domain knowledge. Besides, no dominant feature representation has emerged yet in the literature. In recent years, there has been a growing interest for learning features directly from raw data in the field of audio or video content analysis. Within this context, deep learning methods such as convolutional neural networks (CNNs) and deep belief networks are shown to provide promising results (e.g., [10, 15]). The advantages of deep learning architectures are: (1) *feature re-use*: constructing multiple levels of representation or learning a hierarchy of features, and growing ways to re-use different parts of a deep architecture by changing the depth of the architecture; (2) *abstraction and invariance*: more abstract concepts often can be constructed in terms of less abstract ones and have potentially greater predictive power (i.e., less sensitive to changes in input data) [2].

Inspired by the recent success of the deep learning methods in the field of audio-visual content analysis (e.g., [10,12,15]), we propose to directly learn feature representations from automatically extracted low-level audio-visual features by deep learning for the task of video affective content analysis. The aim of this work is to investigate the discriminative power of mid-level audio-visual representations which are learned from raw data by CNNs for modeling affective content of videos.

Our approach differs from the existing works (presented in Section 2) in the following aspects: (1) we learn both audio and visual feature representations from automatically extracted raw data by using a CNN and fuse these representations at decision-level for the affective classification of music video clips by SVM; (2) we show that the learned mid-level audio-visual representations are more discriminative and provide more precise results than low-level audio-visual ones.

The paper is organized as follows. Section 2 explores the recent developments and reviews methods which have been proposed for affective content analysis of video material with an emphasis on the feature representation of videos. In Section 3, we introduce our method for the affective classification of music video clips. We provide and discuss evaluation results on a subset of the DEAP dataset [11] in Section 4. Finally, we present concluding remarks and future directions to expand our method in Section 5.

2 Related Work

The most common type of affective content analysis approaches is to employ low-level audio-visual features as video representations. In [7], Hanjalic et al. utilize motion, color and audio features to represent arousal and valence. Soleymani et al. [16] address the affective representation of movie scenes based on the emotions that are actually felt

by the audience, where audio-visual features as well as physiological responses of participants are employed to estimate the arousal and valence degree of scenes. Canini et al. [4] aim at defining the emotional identity of a movie in a multidimensional space (along natural, temporal and energetic dimensions) based on audio-visual features in order to retrieve emotionally similar movies based on their emotional identities. In [6], a method for mood-based classification of TV Programs on a large-scale dataset is presented, in which frame-level audio-visual features are used as video representations. Srivastava et al. [17] address the recognition of emotions of movie characters. Low-level visual features based on facial feature points are employed for the facial expression recognition part of the work, whereas lexical analysis of dialogs is performed in order to provide complementary information for the final decision. Cui et al. [5] address affective content analysis of music videos, where they employ audio-visual features for the construction of arousal and valence models. Intended emotion tracking of movies is a subject addressed by Malandrakis et al. [13], where audio-visual features are extracted for the affective representation of movies. In [19], a combined analysis of low-level audio- and visual representations based on early feature fusion is presented for the facial emotion recognition in videos. Yazdani et al. [23] present a method which employs audio-visual features as representation for the affective analysis of music video clips.

The methods of the second category are based on mid-level or hierarchical representations of videos. These solutions construct mid-level representations based on low-level ones and employ these mid-level representations for the affective content analysis of videos. The work presented in [21] combines both low-level audio-visual representations with higher-level video representations. In a first step, movies of different genres are clustered into different arousal intensities (i.e., high, medium, low) with fuzzy c-means [3] using low-level audio-visual features of video shots. In a second step, the results from the first step (i.e., higher level video representations) are employed along together with low-level audio-visual features in order to perform emotional movie classification. Irie et al. [8] propose a latent topic driving model (LTDM) in order to address the issue of classifying movie scenes into affective categories at video shot level. For emotion modeling, the authors adopt Plutchik's eight basic emotions [14] and add a "neutral" category in order to reject movie scenes that arouse no emotion. Video shots are represented with a histogram of quantized audio-visual features and emotion topics are subsequently extracted via latent Dirichlet allocation. Emotions contained in a movie shot are determined based on the topics of the movie shot and predefined emotion transition weights based on the Plutchik's emotion theory. In an extended version of this work [9], Irie et al. propose to represent movie shots with so-called Bag-of-Affective Audio-visual Words and then apply the same LTDM architecture. Xu et al. [22] present a three-level affective content analysis framework, in which the purpose is to detect the affective content of videos (i.e., horror scenes for horror movies, laughable sections for sitcoms and emotional tagging of movies). They introduce mid-level representations which indicate dialog, audio emotional events (i.e., horror sound and laughter) and textual concepts (i.e., informative emotion keywords).

The abovementioned works represent videos with low- or mid-level hand-crafted features. However, in attempts to extend the applicability of methods, there is a growing interest for directly and automatically learning features from extracted low-level

(i.e., raw) audio-visual features rather than representing audio or video data based on manually designed features. Schmidt et al. [15] address the feature representation issue for automatic detection of emotions in music by employing regression-based deep belief networks to directly learn features from magnitude spectra instead of manually designed feature representations. By taking into account the dynamic nature of music, they also investigate the effect of combining multiple timescales of aggregated magnitude spectra as a basis for feature learning. These learned features are then evaluated in the context of multiple linear regression. Li et al. [12] propose to perform feature learning for music genre classification and use CNNs for the extraction of musical pattern features. Ji et al. [10] address the automated recognition of human actions in surveillance videos and develop a novel 3D CNN model for action recognition. The proposed model extracts features from both spatial and temporal dimensions by performing 3D convolutions to capture motion information encoded in multiple adjacent frames. Information coming from multiple channels is combined into a final feature representation. They propose regularizing the outputs with high-level features and combining the predictions of a variety of different CNN models. The proposed method is tested on the TRECVID surveillance video dataset and has proven to achieve superior performance in comparison to baseline methods. Different from the aforementioned existing works, we learn both audio and visual feature representations by using a CNN and perform the affective classification of music video clips by fusing these representations at decision-level by a couple of SVMs. In this work, it is also experimentally shown that the learned audio-visual representations are more discriminative than low-level audio-visual ones.

3 The Video Affective Analysis Method

In this section, we present our approach for the affective classification of music video clips into the four quadrants of the VA-space. An overview of our method is illustrated in Figure 1. Music video clips are first segmented into pieces, each piece lasting 5 seconds, and subsequently MFCC and color values in the RGB space are extracted. After this feature extraction, the next step in the training phase is to build mid-level audio and visual representation generators using CNNs. Eventually, mid-level audio and visual representations are constructed and classifiers are generated using these representations. In the test phase, audio and visual representations are created by using the mid-level representation generators. These representations are used in order to classify a video segment of 5-second length into one of the four quadrants in the VA-space.

The audio and visual feature learning phases of our method are discussed in detail in Section 3.1 and 3.2, respectively. The generation of an affective analysis model is discussed in more detail in Section 3.3.

3.1 Learning Audio Representations

We extract MFCC values for each video segment. The resulting MFCC feature vectors are given as input to a CNN. The first layer (i.e., the input layer) of the CNN is a 125x13 map which contains the MFCC feature vectors from 125 frames of one music video segment. In Figure 2, the CNN architecture used to generate audio representations

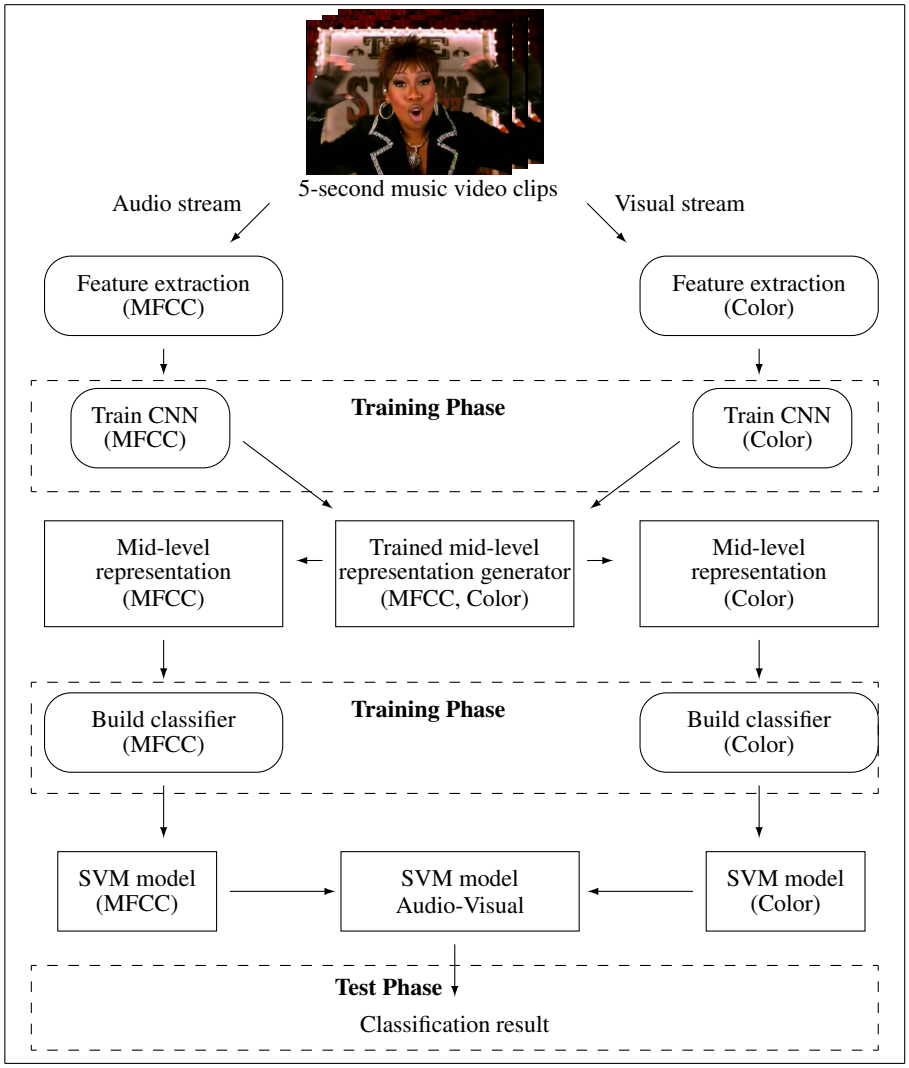


Fig. 1. A high-level overview of our method for video affective analysis. Final classification decisions are realized by a majority voting process. (CNN: Convolutional Neural Network, MFCC: Mel-Frequency Cepstral Coefficients, SVM: Support Vector Machine)

is presented. The CNN has three convolutional and two subsampling layers, one full connection and one output layer (this network size in terms of convolution and subsampling layers has experimentally given satisfactory results). The CNN is trained using the backpropagation algorithm. After training the CNN, the output of the last convolutional layer is used as the mid-level audio representation of corresponding video segment. Hence, the MFCC feature vectors from 125 frames of one segment are converted into a

348-dimensional feature vector (which constitutes a more abstract audio representation) capturing the acoustic information in the audio signal of the music video segment.

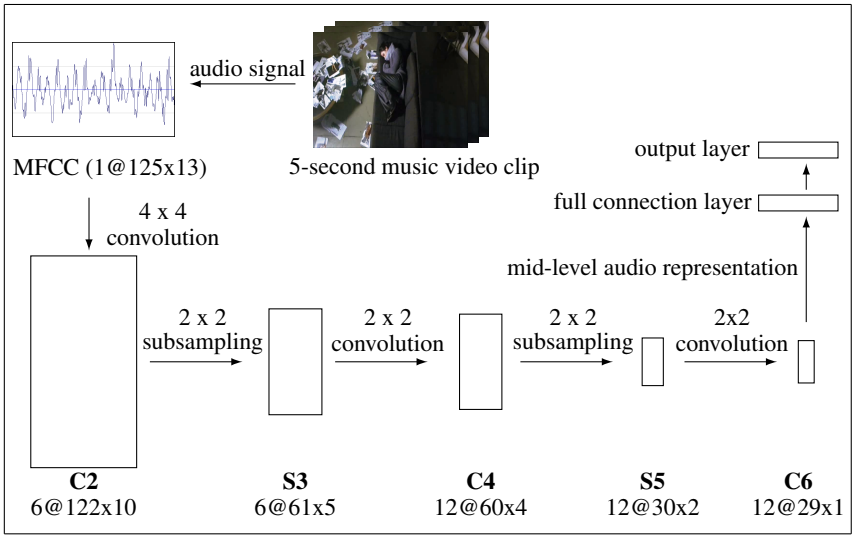


Fig. 2. A CNN architecture for audio affective content analysis. The architecture contains three convolution and two subsampling layers, one full connection and one output layer. (CNN: Convolutional Neural Network, MFCC: Mel-Frequency Cepstral Coefficients)

3.2 Learning Visual Representations

Existing works (e.g., [18]) have shown that colors and their proportions are important parameters to evoke emotions. This observation has motivated our choice of color values for the generation of visual representations for music videos. Keyframes (i.e., representative frames) are extracted from each music video clip segment, as the frame in the middle of a 5-second video segment. For the generation of mid-level visual representations, we extract color information in the RGB space from the keyframe of each segment. The resulting color values in each channel are given as input to a separate CNN. In Figure 3, the CNN architecture used to generate visual representations is presented. The first layer (i.e., the input layer) of the CNN is a 160x120 map which contains the color values from one color channel of the keyframe. The CNN has three convolutional and two subsampling layers, one full connection and one output layer (this network size in terms of convolution and subsampling layers has experimentally given satisfactory results). The training of the CNN is done similarly to the training of the CNN in the audio case. As a result, the color values in each color channel are converted into an 88-dimensional feature vector. The feature vectors generated for each of the three color channels are concatenated into a 264-dimensional feature vector which forms a more

abstract visual representation capturing the color information in the keyframe of the music video segment.

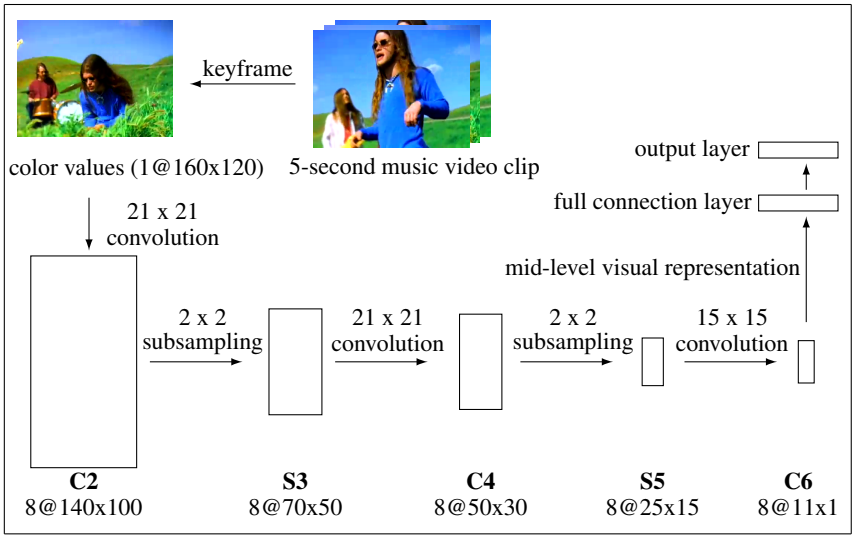


Fig. 3. A CNN architecture for visual affective content analysis. The architecture contains three convolution and two subsampling layers, one full connection and one output layer. (CNN: Convolutional Neural Network)

3.3 Generating the Affective Analysis Model

In order to generate an audio affective analysis model, mid-level audio representations are fed into a multi-class SVM. Similarly, a visual affective analysis model is also generated by feeding mid-level visual representations into a second multi-class SVM. The probability estimates of the two SVM models are subsequently fed into a third multi-class SVM to generate an audio-visual affective video analysis model. Normally, in a basic SVM, only class labels or scores are output. The class label results from thresholding the score, which is not a probability measure. The scores output by the SVM are converted into probability estimates using the method explained in [20].

In the test phase, mid-level audio and visual representations are created by using the corresponding CNN models for music video segments of 5-second length based on MFCC feature vectors and color values in the RGB color space. The music video segments are then classified by using the affective video analysis models. Final decisions for the classification of music video segments are realized by a majority voting process as in [10].

4 Performance Evaluation

The experiments presented in this section aim at comparing the discriminative power of mid-level audio-visual representations against low-level audio-visual features. A direct comparison with the methods (e.g., [23]) which are also tested on the DEAP dataset is limited due to the usage of different subsets of the DEAP dataset (e.g., in [23], only a subset of 40 video clips from the DEAP dataset form the basis of the experiments). An overview of the DEAP dataset is provided in Section 4.1. In Section 4.2, we present the experimental setup. Finally, we provide results and discussions in Section 4.3.

4.1 Dataset and Groundtruth

The DEAP dataset is a dataset for the analysis of human affective states using electroencephalogram, physiological and video signals. We have used all the music video clips whose YouTube links are provided in the DEAP dataset and that were available on YouTube at the time when experiments were conducted (76 music clips). These 76 videos of different affective categories downloaded from YouTube equate to 3,508 music video segments of 5-second length. The dataset is divided into a training set consisting of 2,605 segments from 57 music video clips and a test set consisting of 903 segments from the remaining 19 music video clips.

In the DEAP dataset, arousal and valence values of music clips are in the range of 1 to 9. The arousal values range from *calm / bored* to *stimulated / excited*, while for the valence values the range is from *unhappy / sad* to *happy / joyful*. We have four affective labels used for classification. These are *negative-high*, *negative-low*, *positive-high* and *positive-low* each representing one quadrant in the VA-space. The online ratings of music video clips provided within the DEAP dataset are used in order to determine the label of the music video clips. First of all, the average of the online ratings of the music video clips is computed both for the arousal and valence dimensions. If the average arousal value of a music video clip is above / below 5, then the music video clip is labeled as *high* / *low*. Similarly, a music video clip is labeled as *positive* / *negative*, when the average valence value of the music video clip is above / below 5. Table 1 summarizes the main characteristics of the dataset in more detail.

Table 1. The characteristics of training and test datasets (NH: *negative-high*, NL: *negative-low*, PH: *positive-high*, PL: *positive-low*)

Dataset	**Clips**	**Segments**	**NH**	**NL**	**PH**	**PL**
Train	57	2,605	12	14	12	19
Test	19	903	4	5	4	6
Whole	76	3,508	16	19	16	25

4.2 Experimental Setup

We employed the MIR Toolbox v1.4[1] to extract the 13-dimensional MFCC features. Frame sizes of 40 ms without overlap are used to temporally match with the 25-fps video frames. Mean and standard deviation for each dimension of the MFCC feature vectors are computed, which compose the low-level audio representations of music video segments. In order to generate the low-level visual features of music video segments, we constructed 256-bin color histograms for each color channel in the RGB color space resulting 768-dimensional low-level visual feature vectors. We used the Deep Learning toolbox[2] in order to generate mid-level audio and visual representations with a CNN. We trained the multi-class SVMs with an RBF kernel using libsvm[3] as the SVM implementation. Training was performed using audio and visual features extracted at the music video segment level. More specifically, we trained one SVM using the CNN-based mid-level audio features, one SVM using the CNN-based mid-level visual features and another two SVMs using the low-level audio and visual features as input, respectively. Fusion of audio and visual features is performed at decision-level using an SVM both for low-level and mid-level audio-visual representations. The SVM parameters were optimized by 5-fold cross-validation on the training data. Our approach was evaluated using a training-test split (75% training and 25% test).

4.3 Results and Discussions

Table 2 reports the classification accuracies of our method compared to the method which employs low-level audio-visual features on the DEAP dataset. We achieved 52.63% classification accuracy with mid-level audio-visual representations learned from raw data, while the classification accuracy is 36.84% for the method which uses low-level audio-visual features. When CNN-generated representations are used for training, trained classifiers are able to discriminate better between videos with varying affective content. This demonstrates the potential of our approach for video affective content analysis.

Table 2. Classification accuracies on the DEAP dataset (with audio-visual representations)

Method	**Accuracy (%)**
Our method (mid-level audio-visual)	**52.63**
The low-level audio-visual method	36.84

We present the classification accuracies in the case where only one modality is employed in Table 3. It gives an estimation of the influence in the performance of each modality (either audio or visual) in more detail.

One significant point which can be inferred from Table 3 is that audio representations are more discriminative than visual features for the affective analysis of music

[1] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[2] https://github.com/rasmusbergpalm/DeepLearnToolbox/

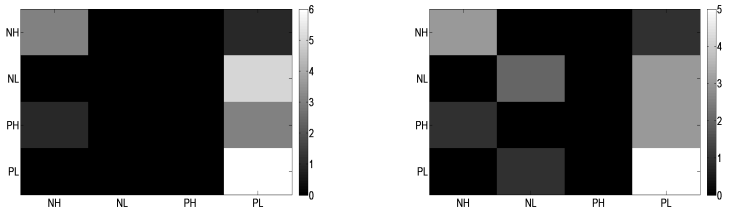
[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Table 3. Classification accuracies on the DEAP dataset (with unimodal representations (audio- or visual-only))

Method	**Accuracy (%)**
Our method (mid-level audio)	**47.37**
Our method (mid-level visual)	36.84
The low-level audio method	36.84
The low-level visual method	31.58

videos. This is true for the mid-level audio representation outperforming the mid-level visual one, and also for the low-level audio representation compared to the low-level visual representation. Additionally, we see that the improvement of using mid-level audio features (around 10% gain compared to low-level audio) compared to the improvement of using mid-level visual features (around 5% gain compared to low-level visual) is higher.

Another conclusion that can be derived from the results is that mid-level audio and visual representations outperform low-level audio and visual representations, respectively. The most significant point which can be deduced from the overall results is that the fusion of mid-level audio and visual representations further improves the performance, while the fusion of low-level audio and visual features provides no significant performance improvement over low-level audio representations. In other words, using color histograms provides a slight improvement (at music segment level) for the affective analysis of music videos. In Figure 4, the confusion matrices of the classification



(a) (b)

Fig. 4. Confusion matrix on the DEAP dataset with (a) mid-level audio and (b) mid-level audio-visual representations learned from raw data (Mean accuracy: 47.37 for audio-only and 52.63 for audio-visual). (NH: *negative-high*, NL: *negative-low*, PH: *positive-high*, PL: *positive-low*).

results of our method (with audio-only and audio-visual representations) for the DEAP dataset are illustrated. The confusion matrix on the left (Figure 4(a)) represents the performance of our method with CNN-generated audio representations, while the confusion matrix on the right (Figure 4(b)) represents the performance of our method with CNN-generated audio and visual representations that are fused at decision-level using a multi-class SVM. The detailed definition of the labels presented in Figure 4 is given in Section 4.1. It is observed from these two matrices that more discriminative video

representations are built by incorporating mid-level color information and that video clips which were wrongly classified as *positive-low* with audio-only representations are classified correctly as *negative-low*. As illustrated in Figure 4(b), *negative-low* - *positive-low* pairs are the most confused affective label pairs. Another point to mention is that our method shows difficulties in discriminating video segments of *positive-high* and classifies these segments mainly as *positive-low*. These results suggest that there is a need to incorporate additional audio and visual features for the representation of video segments.

5 Conclusions

In this paper, we presented an approach for the affective labeling of music video clips, where higher level representations were learned from low-level audio-visual features using CNNs. More specifically, MFCC was employed as audio features, while color values in the RGB space formed the visual features. We utilized the learned audio-visual representations to classify each music video clip into one of the four quadrants of the VA-space using multi-class SVMs. Experimental results on a subset of the DEAP dataset support our belief that higher level audio-visual representations which are learned using CNNs are more discriminative than low-level ones. The current method only exploits MFCC and color features for the generation of video representations. An interesting research question would consist in determining whether augmenting the feature set by including an extensive set of audio and visual features (especially motion-related features such as optical flow) would provide a significant gain in performance. Hence, as future work, we plan to study the representation of videos with additional audio- and visual features. Exploring the architecture of the CNNs (i.e., adjusting the number of convolution layers and kernel size) is another direction for future research. Another direction in the context of learning features by deep learning is to use unsupervised deep learning methods (e.g., restricted Boltzmann machines or stacked auto-encoders [1]) for modeling the data before applying SVM. Comparing the performance of learned audio-visual representations to hand-crafted mid-level feature representations such as the bag-of-words approaches is another track for future work.

References

1. Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
2. Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. 2013.
3. J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.
4. L. Canini, S. Benini, P. Migliorati, and R. Leonardi. Emotional identity of movies. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1821–1824. IEEE, 2009.
5. Y. Cui, J. S. Jin, S. Zhang, S. Luo, and Q. Tian. Music video affective understanding using feature importance analysis. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 213–219. ACM, 2010.

6. J. Eggink and D. Bland. A large scale experiment for mood-based classification of tv programmes. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 140–145. IEEE, 2012.
7. A. Hanjalic and L. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, pages 143–154, 2005.
8. G. Irie, K. Hidaka, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Latent topic driving model for movie affective scene classification. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 565–568. ACM, 2009.
9. G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa. Affective audio-visual words and latent topic driving model for realizing movie affective scene classification. *Multimedia, IEEE Transactions on*, 12(6):523 –535, oct. 2010.
10. S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 221–231, 2013.
11. S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, 2012.
12. T. Li, A. B. Chan, and A. Chun. Automatic musical pattern feature extraction using convolutional neural network. In *Proc. Int. Conf. Data Mining and Applications*, 2010.
13. N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi. A supervised approach to movie emotion tracking. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2376–2379. IEEE, 2011.
14. R. Plutchik. The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.
15. E. M. Schmidt, J. Scott, and Y. E. Kim. Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *International Society for Music Information Retrieval*, pages 325–330, 2012.
16. M. Soleymani, G. Chanel, J. Kierkels, and T. Pun. Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 228–235. IEEE, 2008.
17. R. Srivastava, S. Yan, T. Sim, and S. Roy. Recognizing emotions of characters in movies. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 993–996. IEEE, 2012.
18. P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394, 1994.
19. M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig. Low-level fusion of audio and video feature for multi-modal emotion recognition. In *3rd International Conference on Computer Vision Theory and Applications. VISAPP*, volume 2, pages 145–151, 2008.
20. T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, 5:975–1005, 2004.
21. M. Xu, J. S. Jin, S. Luo, and L. Duan. Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 677–680. ACM, 2008.
22. M. Xu, J. Wang, X. He, J. S. Jin, S. Luo, and H. Lu. A three-level framework for affective content analysis and its case studies. *Multimedia Tools and Applications*, 2012.
23. A. Yazdani, K. Kappeler, and T. Ebrahimi. Affective content analysis of music video clips. In *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12. ACM, 2011.